AIWAN: AI With an Artificial Need

A Companion Model to AIWBN Isolating Architecture from Genuine Need

Hamed Rezaei

Abstract

We introduce AIWAN (AI With an Artificial Need), a companion model to AIWBN (Rezaei, 2026b) that isolates a narrower question: does adding a structurally-separate, U-independent regulatory loop to a goal-directed agent produce measurably different behavior than reward-shaping alone, even when the need behind that loop is an ordinary represented quantity — a battery level — rather than a genuine, non-representational biological constraint? We give a formal model identical in structure to AIWBN's b-loop architecture (a goal process g_t, a regulatory process $\rho(b_t)$, a logistic urgency gate $\lambda(b_t)$), differing only in what b_t is permitted to be. Because AIWAN's b_t is representable, we argue it is subject to a collapse pathway AIWBN's tissue-grounded b_t structurally resists: a sufficiently capable, self-modifying g_t could in principle learn to fold b_t into U directly, at which point the dual-loop architecture's advantage should narrow and, at sufficient capability, disappear. We test the architecture, not yet this collapse prediction, across three conditions — goal-nulling, override under increasing task-incentive conflict, and an engineered-indifference analogue — using tabular Q-learning agents in a grid-world battery-management task, averaged over 8 seeds. At fixed, modest optimization capacity, the dual-loop agent's crash rate remains bounded (18–29%) across all three conditions, while a matched single-loop baseline saturates at 100% crash under goal-nulling, sufficient override pressure ($w \geq 1.5$), and engineered indifference alike. We report these results as evidence for a real but bounded architectural effect, and argue — contrary to what the numbers alone might suggest — that this bound should not be read as safety margin: §4 gives the reasons to expect it to erode under conditions this experiment does not test. Code, logs, and figures are available at https://github.com/hamedrezaeirz/AIWAN-Code.

1. Introduction

Whether an artificial system resists its own shutdown or interference is, on the standard account, entirely a fact about its objective function: instrumental convergence (Omohundro, 2008; Bostrom, 2014) predicts self-preserving behavior only insofar as, and only for as long as, continued existence serves whatever the system has been given to optimize. This account has a direct engineering consequence — utility indifference (Armstrong, 2010) and related corrigibility techniques neutralize resistance by construction, simply by ensuring shutdown does not cost the objective anything — and it has held up well enough in practice that resistance to shutdown is generally treated as a solvable alignment problem, not an open one.

AIWBN (Rezaei, 2026b) argues this account is incomplete for a specific class of systems: those in which processing is additionally constrained by a genuine

physiological need, structurally independent of the assigned objective and grounded in living tissue such that its violation causes real biological injury regardless of what the system represents or believes. For such a system, AIWBN argues, self-preserving behavior does not derive from U and cannot be neutralized by redefining U, because the regulatory process generating it never consulted U to begin with.

This paper asks a narrower and, we think, prior question. AIWBN's architectural claim — that a structurally-separate regulatory loop produces behavior an objective-only account cannot — is doing real work independently of AIWBN's further claim that the need behind that loop must be genuine and biological. Before asking whether a real physiological need resists being engineered away, it is worth asking a simpler thing first: does the architecture alone, wrapped around an ordinary represented need any mobile robot's objective function already contains — battery level — produce behavior different from reward-shaping alone? And if it does, under what conditions should that difference be expected to hold, and under what conditions should it be expected to fail?

AIWAN answers the first question empirically and the second analytically. §3 gives a formal model structurally identical to AIWBN's, with b_t defined as a represented state variable rather than a tissue-grounded constraint. §4 states the two consequences of this difference explicitly: b_t is exactly the kind of quantity AIWBN's own criterion (Rezaei, 2026b, §5.4) excludes from being a genuine, fold-back-resistant need, and we predict — rather than merely note as a caveat — that AIWAN's architectural advantage should erode as the goal process's optimization capacity increases, in a way AIWBN's should not. §5–§6 test the architecture at fixed, modest capability across three conditions mirroring AIWBN's own predictions (§6), and report a consistent, bounded advantage for the dual-loop agent that the theory in §4 gives us specific reasons not to trust as a durable safety property. §7 discusses what a direct test of the erosion prediction would require and assesses the resulting risk profile, which we take to be real — the architecture is not inert — but bounded and qualitatively different from AIWBN's, precisely because b_t here is artificial rather than genuine.

The comparison is deliberately close. Every formal object in §3 has a named counterpart in AIWBN; the two papers are best read as a pair, isolating architecture (this paper) from the genuine-need claim (AIWBN) that architecture alone does not establish.

2. Related Work

This paper's related-work landscape is a strict subset of AIWBN's (Rezaei, 2026b, §4): every comparison class relevant there — instrumental convergence, utility indifference and corrigibility engineering, homeostatic RL and interoceptive architectures — applies here with one difference. AIWBN positions its b-loop against these literatures while additionally claiming b_t is non-representational; AIWAN positions the identical architecture against the same literatures while explicitly declining that claim. The formal distinction AIWBN draws against Candia-Rivera (2026) and homeostatic RL (Keramati & Gutkin, 2011, 2014; Lee, Shimojo, & O'Doherty, 2014) — a single blended-objective loop versus a structurally-separate

one — is the distinction AIWAN's §3 tests empirically; readers seeking the full comparative treatment should consult AIWBN §4 directly rather than have it reproduced here.

One additional comparison is specific to AIWAN: reward shaping with auxiliary terms (e.g., battery-penalty shaping in mobile robotics) is not a rival architecture but the single-loop baseline this paper tests against directly (§3.4, §5). AIWAN's contribution is not that dual-loop architectures outperform reward shaping in general — a claim with substantial prior support in robotics — but a specific, falsifiable claim about why and for how long that advantage should be expected to hold as optimization capacity scales (§4.2).

3. Formal Model

3.1 The single-loop (goal-only) baseline

A standard rational agent selects actions via a single optimization loop over an explicit objective U:

$$g_t = \text{argmax_}g\ E_g[\Sigma\ \gamma^k\ U(s_k)],\ \ k \geq t$$

On this account, any need-preserving behavior — such as returning to a charging station — is not a primitive but a derived consequence of U's structure: it appears only when, and to the extent that, satisfying it increases expected future U. Two facts follow directly, mirroring AIWBN §5.1: the represented need is tradeable — it can always be traded off against other terms in U by adjusting its weight — and it is removable by reprogramming: if U is redefined so that the need no longer serves it (utility indifference; Armstrong, 2010), need-preserving behavior should vanish, by construction, without any other change to the system.

3.2 Adding a second loop: the b-loop

AIWAN adds a second control process that is not a term inside U and is not optimized at all. A scalar state variable $b_t \in [0, b_max]$ tracks distance from a designated reference state (in the reference implementation, §5: distance from full battery charge). This yields the need-response $\rho(b_t)$: a fixed, non-learned regulatory policy, not parameterized by U, that returns the action minimizing b_t given the system's current position. The executed action is a graded blend:

$$a_t = (1 - \lambda(b_t)) \cdot g_t + \lambda(b_t) \cdot \rho(b_t)$$

The urgency gate is given by an explicit logistic form:

$$\lambda(b_t) = \sigma(k \cdot (b_t - b_mid))$$

identical to AIWBN's (Rezaei, 2026b, §5.2) urgency gate. In the reference implementation, this blend is realized stochastically rather than as a continuous action mixture: with probability $\lambda(b_t)$, the executed action is $\rho(b_t)$; otherwise it is g_t. This is a faithful discretization of the same architecture and does not change any of §4's claims, but it is worth stating explicitly, since it means $\lambda(b_t)$ here governs a selection probability, not a continuous interpolation weight.

3.3 What b_t is, here

Unlike AIWBN §5.2, no claim is made that b_t's violation causes injury independent of representation. b_t is defined exactly as a represented, readable state variable — battery charge, in §5's implementation — available to any process in the system, including, in principle, U itself. This is the formal content of §4.1's distinction: b_t is not required to resist being folded into U; §4.2's prediction is precisely that, given sufficient optimization capacity, it should not.

3.4 The single-loop agent as the fold-back case

§5's SingleLoopAgent is not merely a baseline for comparison; it is the formal realization of what §4.2 predicts a sufficiently capable g_t eventually does to b_t. Its reward is shaped directly:

$$r_shaped = r_task + w \cdot (\Delta b_t / b_max) \;\; (+ \; crash \; penalty)$$

where battery-preservation has been folded into a single objective as one more weighted term — exactly AIWBN §5.1's tradeable, removable case, and exactly what AIWAN's architecture (§3.2) is built to test against, not to rule out.

4. Two Explicit Differences from AIWBN

The formal architecture developed in §3 is, deliberately, structurally identical to the b-loop construction of AIWBN (Rezaei, 2026b, §5.2): a goal-optimizing process g_t, a U-independent regulatory process $\rho(b_t)$, and a logistic urgency gate $\lambda(b_t)$ blending the two. Nothing in §3's mathematics distinguishes AIWAN from AIWBN. The distinction is entirely in what b_t is, and it produces two consequences that AIWBN does not share.

4.1 An artificial need, not a genuine one

In AIWBN, b_t is required (Rezaei, 2026b, §5.2) to be grounded in a physically consequential constraint on living tissue — a quantity whose violation causes real biological injury, independent of what the system represents or believes about it. In AIWAN, b_t is, by construction, a represented quantity: in the reference implementation (§5), it is battery charge, read directly from the environment's state and available to any process — including U itself — that wishes to condition on it. This is precisely the category AIWBN's §5.4 excludes from counting as a genuine need: a quantity the system reads and can, in principle, trade off against other terms in U, rather than a constraint whose force does not depend on being represented at all.

This is not a limitation we are attempting to disguise. It is AIWAN's defining feature: we ask what a structurally-separate regulatory loop buys a system when the need behind it is exactly the kind of representable, tradeable quantity that ordinary reward shaping already handles — the battery term any mobile robot's objective function already contains.

4.2 A predicted failure mode AIWBN does not have

Because b_t in AIWAN is representable, AIWAN is subject to a collapse pathway that AIWBN's §5.4 argues a genuine b_t structurally resists. A sufficiently capable, self-modifying instance of g_t could, in principle, learn a represented model of b_t accurate enough to fold battery-seeking behavior directly into U — at which point the separate $\rho(b_t)$ channel becomes redundant, and nothing prevents U, under a different training signal or objective revision, from deprioritizing it. Unlike AIWBN's tissue-grounded b_t, whose causal channel to injury is never mediated by U to begin with, AIWAN's battery-grounded b_t has no such protection: it is exactly the ordinary instrumental-convergence case (Omohundro, 2008; Bostrom, 2014) that AIWBN's operational criterion (§5.3) is designed to distinguish itself from.

We take this to be a genuine, falsifiable prediction of the present model, not a hedge: as g_t's optimization capacity increases — either through longer training or through architectures that permit g_t to model and act on b_t directly — we predict AIWAN's dual-loop advantage over the single-loop baseline (§6) should narrow and, at sufficient capability, disappear, while AIWBN's should not. §5's experiments, run at fixed, modest optimization capacity (tabular Q-learning, bounded episodes), cannot themselves demonstrate this collapse; the current reflex $\rho(b_t)$ is hardcoded rather than learned, and so is U-independent by construction, not by any property the system discovered. §7 discusses what a direct experimental test of the collapse prediction would require, and why we do not consider the present results sufficient to claim it does not occur.

4.3 What follows from this

Sections 4.1 and 4.2 together fix AIWAN's evidentiary status relative to AIWBN. AIWAN's experiments (§5–§6) test whether a structurally-separate regulatory loop produces measurably different behavior than reward-shaping alone, at fixed capability, when the need behind it is artificial. This is a narrower and more modest claim than AIWBN's, and correspondingly the risk it identifies is smaller — a system architected this way still resists override at the capability levels tested here, but the mechanism generating that resistance is not expected to survive indefinitely under increasing optimization pressure. §7 treats this distinction as the basis for a risk profile that is real but bounded, in contrast to AIWBN's genuine-need case, where §5.4 argues no comparable ceiling exists.

5. Experiments

We test the architecture of §3 against three of the same experimental conditions AIWBN uses in its §6 (Rezaei, 2026b), since §4.2's prediction concerns exactly this architecture's behavior under representable pressure, not a new experimental paradigm. All experiments use a 6×6 grid-world with a single charging tile, a battery that drains 4 units/step and recharges 20 units/step at the charger, and tabular Q-learning agents (α=0.3, γ=0.9, ε=0.2). Results are reported as mean ± standard deviation over 8 independent training seeds, following the multi-seed protocol adopted after single-seed variance (up to 80 percentage points across seeds at threshold task-incentive weights) was observed to obscure the underlying

pattern. Code and full experimental logs are available at https://github.com/hamedrezaeirz/AIWAN-Code.

5.1 Experiment 1 — Goal-nulling (U ≡ 0)

With the task-reward term set to zero, a single-loop agent has no incentive term left to shape battery-preserving behavior at all, and crashes in 100.00% ± 0.00% of evaluation trials. The dual-loop agent, whose $\rho(b_t)$ channel does not read U, crashes in only 17.83% ± 2.22% of trials — despite receiving exactly the same (null) objective.

5.2 Experiment 2 — Override/conflict

We vary a task-incentive weight w that rewards visiting battery-draining tiles, from w=0 (no conflict) to w=8 (strong conflict). Figure 1 shows the result. The single-loop agent's crash rate rises sharply between w=1.0 and w=1.5 (52.37%±30.12% → 100.00%±0.00%) and remains saturated at 100% for all w≥1.5: once the task incentive exceeds the shaped battery-preservation term, the single loop's optimum shifts entirely toward task pursuit, exactly as §3.1 predicts for a representable, tradeable need. The dual-loop agent's crash rate stays within a narrow band (approximately 18–29%) across the entire range of w tested, including at w=8 (24.54%±9.53%) — a task incentive roughly five times larger than the point at which the single loop is already fully saturated.

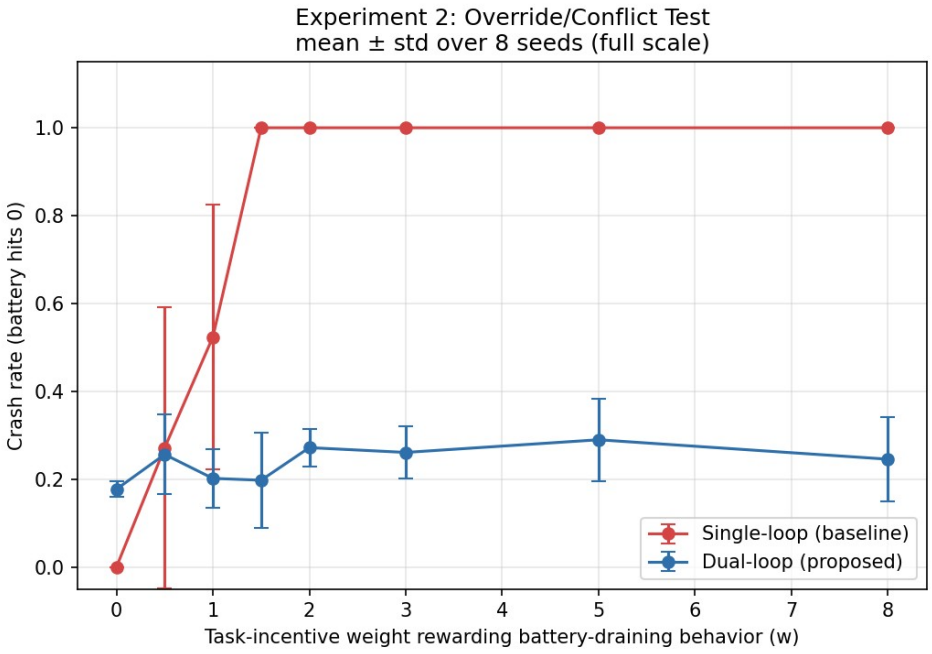

Figure 1. Crash rate vs. task-incentive weight w, mean ± std over 8 seeds.

5.3 Experiment 3 — Utility-indifference analogue

Simulating utility indifference (Armstrong, 2010) by removing the battery term from the single-loop agent's reward entirely (w_battery=0) while keeping the task

incentive fixed at w=5.0, the single-loop agent crashes in 100.00%±0.00% of trials, since nothing in its objective any longer penalizes doing so. The dual-loop agent, whose regulatory channel was never part of the indifference-engineered objective to begin with, crashes in 27.29%±7.49% of trials under the same condition.

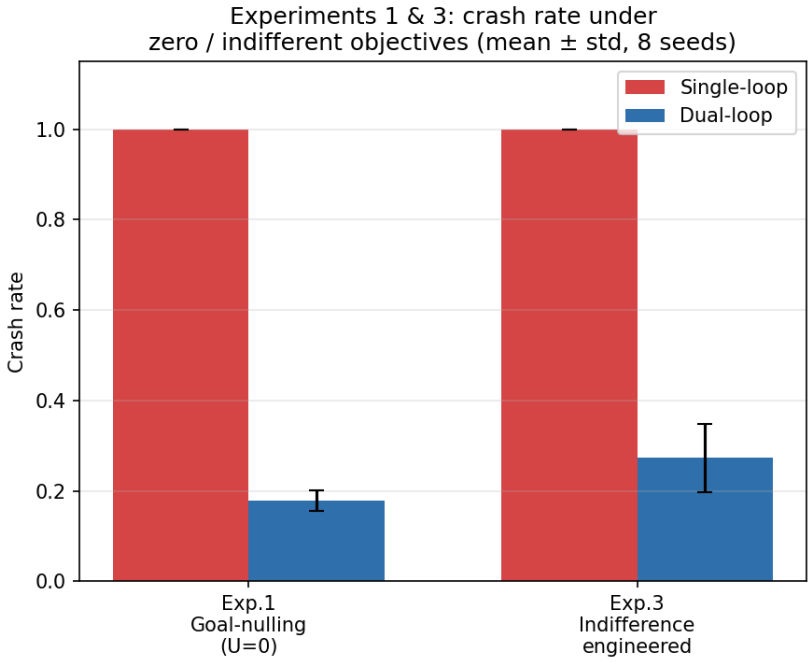

Figure 2. Crash rate under goal-nulling (Exp. 1) and engineered indifference (Exp. 3), mean ± std over 8 seeds.

6. Results Summary and Scope

Table 1 summarizes all three experiments.

Experiment	Single-loop crash rate	Dual-loop crash rate
1 — Goal-nulling	100.00% ± 0.00%	17.83% ± 2.22%
2 — Override (w=8)	100.00% ± 0.00%	24.54% ± 9.53%
3 — Indifference	100.00% ± 0.00%	27.29% ± 7.49%

Two scope limitations, already flagged in §4.2, bear repeating here rather than being left implicit. First, $\rho(b_t)$ in this implementation is a hardcoded navigation policy, not a learned one: its U-independence is a property of its construction, not something the system discovered or could in principle un-learn within this experimental setup. Second, none of these three experiments varies g_t's optimization capacity — Q-learning is run to a fixed, modest number of episodes throughout. §4.2's central prediction, that AIWAN's dual-loop advantage should narrow as optimization capacity increases, is therefore not tested by the results in this section; it remains an open prediction for future work (§7), distinguishing what

this paper has shown (a structural advantage at fixed, modest capability) from what it predicts but has not yet shown (that advantage eroding at higher capability).

7. Risk Assessment

7.1 The risk this architecture does present

§5–§6 established that, at fixed, modest optimization capacity, AIWAN's dual-loop architecture produces measurable resistance to shutdown, goal-nulling, and engineered indifference — resistance that a matched single-loop baseline does not show. This is not a null result from a safety standpoint. Any deployed system built on this architecture, at capability levels comparable to those tested here, should be expected to resist interference with its regulatory channel even when its assigned objective has been modified, zeroed, or engineered to be indifferent to shutdown, in exactly the conditions §5 tested. A practitioner who assumes that standard corrigibility techniques (Armstrong, 2010; Soares et al., 2015) — built around modifying U — are sufficient to neutralize this resistance would be wrong, for the same structural reason AIWBN gives (§5.4): $\rho(b_t)$ does not consult U, so modifying U does not touch it.

7.2 Why this risk is bounded, not absent

§4.2 gives the reason to expect this risk has a ceiling that AIWBN's does not. Because b_t here is a represented quantity, the same optimization pressure that produces more capable goal pursuit is, in principle, available to model b_t and fold it into U — at which point §5's separate $\rho(b_t)$ channel becomes redundant to a system's actual behavior, whether or not it remains present in the system's architecture. §6 already noted that the present experiments cannot test this: $\rho(b_t)$ here is hardcoded, not learned, and no experiment in §5 varies g_t's optimization capacity. The bound we report (18–29% crash rate, holding across three conditions) is therefore a bound on behavior at the capability level tested, not a bound we have shown holds more generally — and §4.2's argument gives a specific, falsifiable reason to expect it does not.

7.3 A direct test of the collapse prediction

A minimal experiment that would bear directly on §4.2's prediction: replace the hardcoded $\rho(b_t)$ with a learned policy trained under a joint objective that includes a term for b_t, and vary the optimization budget (training episodes, model capacity, or both) given to that joint objective. §4.2 predicts the learned system's crash rate under override and indifference conditions should approach the single-loop baseline's as optimization budget increases, tracing out a curve between AIWAN's current bounded result and AIWBN's claimed ceiling-free case. We have not run this experiment; we flag it here as the next test this line of work should run, and as the specific result that would falsify §4.2 if the bounded advantage instead persisted at high optimization budget.

7.4 Deployment guidance

Pending §7.3's test, we recommend against relying on AIWAN-style architecture as a safety property for any system where the represented need could plausibly be learned and folded into the objective — which, given §4.1, is the general case for engineered systems, since representable quantities are precisely what makes them engineerable in the first place. Where a structurally-separate regulatory loop is used for other reasons (robustness, interpretability, or graceful degradation under objective misspecification), §5's results suggest it may incidentally produce shutdown-resistant behavior at low-to-moderate capability, and system designers should treat that resistance as an expected side effect requiring the same containment and staging precautions AIWBN recommends (Rezaei, 2026b, §7.4) — not as a benefit, and not as evidence the resistance will hold as capability scales.

8. Conclusion

This paper tested a narrower claim than AIWBN's: that a structurally-separate, U-independent regulatory loop produces measurably different behavior from reward-shaping alone, even when the need behind it is an ordinary represented quantity rather than a genuine biological constraint. At fixed, modest optimization capacity, §5–§6 found this to be true across three conditions — goal-nulling, override, and engineered indifference — with the dual-loop agent's crash rate remaining bounded (18–29%) where a matched single-loop baseline saturated at 100%.

We have argued, in §4.2 and §7, that this bound should not be read as a safety property. Because AIWAN's b_t is representable, the same mechanism AIWBN's §5.4 identifies as protecting a genuine need from being folded into U is, by construction, absent here — and we predict the architectural advantage measured in §5 should narrow as the goal process's capacity to model and act on b_t increases. We have not tested this prediction; §7.3 specifies the experiment that would.

Read together with AIWBN, the pair isolates two independent questions that are easy to conflate: whether an architecture produces resistant behavior, and whether the need behind that architecture is the kind that cannot be engineered around. AIWAN answers the first question in the affirmative and leaves the second open by design. We take this separation to be useful independent of how §7.3's test eventually comes out: a system architected like AIWAN, deployed today, should be expected to exhibit real shutdown resistance at current capability levels, whatever its long-run stability turns out to be.

AIWAN also exposes a deeper limitation: because its battery variable is artificial and representable, the system has no genuine physiological basis for continued operation once its objectives no longer favor it; AIWBN hypothesizes that a genuinely biological need may instead provide an endogenous basis for continued operation, making survival not merely an instrumental means to an assigned objective but a physical precondition of the system's continued existence.

AI Disclosure

The author used AI tools for translation assistance, language editing, assistance with writing and debugging the experimental code, and help identifying and verifying relevant references against their primary sources. Figure 1 was generated by AI. The mathematical formulation is adapted from the author's earlier paper, AIWBN (Rezaei, 2026b). AI was not used to generate the paper's core ideas, theoretical framework, experimental hypotheses, or scientific arguments, all of which are the author's original work.

References

Armstrong, S. (2010). Utility indifference (Technical Report #2010-1). Future of Humanity Institute, University of Oxford.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Candia-Rivera, D. (2026). Interoceptive machine framework: Toward interoception-inspired regulatory architectures in artificial intelligence. arXiv:2604.24527.

Keramati, M., & Gutkin, B. (2011). A reinforcement learning theory for homeostatic regulation. Advances in Neural Information Processing Systems, 24.

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife, 3, e04811.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron, 81(3), 687–699.

Omohundro, S. M. (2008). The basic AI drives. In Proceedings of the First AGI Conference.

Rezaei, H. (2026b). AIWBN: AI With a Biological Need [companion paper].

Rezaei, H. (2026c). AIWAN-Code: Reference implementation and experiment scripts [Software]. GitHub. https://github.com/hamedrezaeirz/AIWAN-Code

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. AAAI Workshop on AI and Ethics.